Universidade de Brasília
IE – Departamento de Estatística
Estágio Supervisionado 2

Modelos de Regressão Logística em Delineamentos Complexos

Pedro Farage Assunção

Relatório do Projeto Final

Orientador: Prof. Dr. Eduardo Freitas da Silva

Brasília,

2013

Pedro Farage Assunção – 09/0128265

Estágio Supervisionado 2

Modelos de Regressão Logística em Delineamentos Complexos

Orientador: Prof. Dr. Eduardo Freitas da Silva

Trabalho de Conclusão de Curso apresentado à Universidade de Brasília, como exigência parcial à obtenção do título de bacharel em Estatística.

Brasília,

2013

SUMÁRIO

1 Introdução

Este estudo trata de um problema frequentemente encontrado por pesquisadores que obtêm seus dados por meio de delineamentos complexos: como estimar seus parâmetros e obter estimativas confiáveis e adequadas. Silva (2002) define delineamento complexo como: "estratificação das unidades de amostragem, conglomeração (seleção de amostras em vários estágios, com unidades compostas de amostragem), probabilidades desiguais de seleção em um ou mais estágios, e ajustes dos pesos amostrais para calibração com totais populacionais conhecidos".

O principal motivo da preocupação em estimar dados provenientes de amostras complexas diferentemente dos métodos usados quando a amostra aleatória simples é empregada deve-se ao fato de que quando os pesos amostrais são considerados nos cálculos, as estimativas populacionais dos parâmetros são não-viciadas. As estimativas descritivas como a média populacional são influenciadas pelos pesos diferentes das observações e estimativas de variância, desvio padrão e parâmetros de ajuste a alguns modelos são influenciadas tanto pelos pesos das observações quanto pela estratificação e conglomeração utilizadas. Se ignorado estes aspectos de coleta dos dados, as estimativas podem levar a conclusões erradas e inadequadas.

Hoje, pelo avanço e facilidade de uso dos softwares já existem muitos recursos disponíveis para facilitar e melhorar a incorporação adequada dos diversos aspectos amostrais em cada pesquisa, tanto na estimação e precisão dos parâmetros quanto no ajuste, diagnóstico e avaliação de modelos ajustados. Esses pontos levaram a uma melhor interpretação de resultados com maior acurácia e adequabilidade.

O estudo será conduzido por meio de análise de regressão logística englobando os aspectos de planos amostrais complexos. A regressão logística será adequada visto que a aplicação e validação da teoria explicitada no trabalho fará uso de variável resposta dicotômica.

2 Objetivos

2.1 Objetivo Geral

- Estudo de modelos de regressão logística em planos amostrais complexos.

2.2 Objetivos Específicos

- Estudar métodos de estimação e verificar que de acordo com o delineamento utilizado as estimativas devem ser ajustadas;
- Aplicar a metodologia em um banco de dados.

3 Metodologia

3.1 Regressão Logística

Em modelos lineares generalizados a preocupação está no estudo da relação entre a variável resposta e uma ou mais variáveis explicativas. É comum encontrar variáveis respostas discretas assumindo dois ou mais valores e para estes casos a regressão logística é comumente utilizada. A principal diferença da regressão logística para regressão linear é que a variável resposta é binária, que reflete na escolha de modelos paramétricos e suposições. Levando em conta essas diferenças, os métodos usados na análise de regressão logística seguem o mesmo dos empregados na regressão linear.

3.1.1 Modelo

Em qualquer modelo de regressão o interesse está no valor médio da variável resposta dado o valor da variável explicativa, denotado por "E(Y|x)". Tratando-se de regressão logística, variável resposta dicotômica, E(Y|x) é uma proporção e, portanto, $0 \leq E(Y|x) \leq 1$. A curva de E(Y|x) tem forma de S, pois conforme a variável explicativa diminui E(Y|x) gradualmente se aproxima de 0 e quando a variável explicativa aumenta E(Y|x) gradualmente se aproxima de 1.

Para modelagem desse tipo de curva escolhe-se a distribuição logística. Matematicamente é bastante flexível e de fácil uso e leva a interpretações significantes. Usa-se a notação $\pi(x) = E(Y|x)$ para representar a média condicional de Y dado x e ela é definida como:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 * x}}{1 + e^{\beta_0 + \beta_1 * x}} \quad (1)$$

A transformação logito de $\pi(x)$ é um tópico de interesse do estudo. Ela é dada por:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 * x \quad (2)$$

A importância dessa transformação é que $g(x)$ possui propriedades desejáveis de um modelo de regressão linear. O logito, $g(x)$, é linear nos parâmetros, pode ser contínua e, dependendo da alcance de x, pode variar de $-\infty$ a $+\infty$.

Uma observação da variável resposta Y pode ser escrita como $y = \pi(x) + \varepsilon$, onde ε é o erro. Para uma variável resposta dicotômica Y, ε assume um de dois valores possíveis. Se $y = 1,\ \varepsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, e se $y = 0,\ \varepsilon = -\pi(x)$ com probabilidade $1 - \pi(x)$. Disso ε tem distribuição com média 0 e variância $\pi(x) * (1 - \pi(x))$. Portanto, Y|x segue uma Binomial com probabilidade $\pi(x)$.

3.1.2 Estimação dos Parâmetros

Método da Máxima Verossimilhança:

Considere uma amostra de n observações independentes do par $(yi\ ;\ xi)$, i=1, 2, ..., n, onde y_i é o valor da i-ésima variável resposta binária, codificada como 0 ou 1 e x_i o valor da i-ésima variável explicativa. Para ajustar um modelo faz-se necessário estimar os valores de β_o e β_1. Basicamente, o método da máxima verossimilhança obtêm estimadores para $\boldsymbol{\beta} = (\beta_0\ ,\ \beta_1)$ que maximizam a probabilidade de obter os dados observados da amostra.

Definindo a função de verossimilhança como a probabilidade dos dados observados como uma função dos parâmetros $\boldsymbol{\beta}$ temos que: se Y é codificado em 0 e 1, $\pi(x) = P(Y = 1|x)$ e $1 - \pi(x) = P(Y = 0|x)$. Disso segue que quando $yi = 1$ a contribuição para a função de verossimilhança é $\pi(x_i)$ e quando $yi = 0$ a contribuição para a função de verossimilhança é $1 - \pi(x)$. Portanto a contribuição para a função de verossimilhança do par $(xi\ , yi)$ é $\pi(x_i)^{y_i} * [1 - \ \pi(x_i)]^{1-y_i}$.

Como, de pressuposto, as observações são independentes, a função de verossimilhança é obtida pelo produtório das contribuições de cada par $(yi\ ;\ xi)$ indicado acima. Então, tem-se que:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} * [1 - \ \pi(x_i)]^{1-y_i} \quad (3)$$

O objetivo é estimar $\boldsymbol{\beta}$ que maximize a equação acima. Para isso, a manipulação matemática pelo log da verossimilhança é mais fácil. O log da verossimilhança é definido por:

$$l(\boldsymbol{\beta}) = ln\big(L(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \big\{y_i * \ln\big(\pi(x)\big) + (1 - y_i) * \ln\big(1 - \pi(x)\big)\big\} \quad (4)$$

Para maximizar $l(\boldsymbol{\beta})$, deriva-se em relação a β_0 e β_1 e iguala-se a zero o resultado. Com isso, obtêm-se duas equações:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \quad (5)$$

e

$$\sum_{i=1}^{n} x_i * [y_i - \pi(x_i)] = 0 \quad (6)$$

As equações descritas acima são não lineares e, por isso, necessitam de métodos especiais iterativos de estimação. McCullagh e Nelder (1989) mostraram que a solução pode ser obtida usando um processo iterativo de mínimos quadrados ponderados.

Os valores de $\boldsymbol{\beta}$ das equações acima são os estimadores de máxima verossimilhança, $\widehat{\boldsymbol{\beta}}$.

3.1.3 Teste de Significância do Estimador

Depois de estimado $\boldsymbol{\beta}$, é de interesse saber se a variável que teve o β estimado é relevante ou não na análise, ou seja, se o modelo com a variável explicativa em questão explica a variável resposta melhor que o do modelo sem a variável explicativa.

Os métodos em regressão logística seguem o mesmo princípio que em regressão linear: comparar os valores observados com os valores preditos da variável resposta. Essa comparação é baseada na função de verossimilhança e é amplamente conhecida como teste da razão de verossimilhança. Ela é baseada na seguinte função:

$$D = -2 * \ln\left[\frac{verossimilhança\ do\ modelo\ ajustado}{verossimilhança\ do\ modelo\ saturado}\right]$$

$$= -2 * \sum_{i=1}^{n}\left[y_i * \ln\left(\frac{\widehat{\pi_i}}{y_i}\right) + (1 - y_i) * \ln\left(\frac{1 - \widehat{\pi_i}}{1 - y_i}\right)\right] \quad (7)$$

Onde $\widehat{\pi_i} = \hat{\pi}(x_i)$.

Para avaliar a significância de uma variável explicativa, comparam-se os valores de D com e sem a variável em questão e verifica se é significativo. Para isso usa-se:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável)$$
$$= -2 * \ln\left[\frac{verossimilhança\ sem\ a\ variável}{verossimilhança\ com\ a\ variável}\right] \quad (8)$$

Sob H_0: β_1=0, G segue uma Qui-Quadrado com 1 grau de liberdade.

Teste de Wald:

O teste de Wald é obtido pela comparação do estimador de máxima verossimilhança $\hat{\beta}_1$ com a estimação de seu erro.

$$W = \frac{\widehat{\beta_1}}{\widehat{SE}(\widehat{\beta_1})} \quad (9)$$

Onde $\widehat{SE}(\widehat{\beta 1})$ é a estimativa do erro padrão do parâmetro estimado. Sob H_0: β_1=0, W segue uma normal padrão.

Score Test:

Tanto o teste da razão de verossimilhança quanto o teste de Wald requerem o cálculo computacional do estimador de máxima verossimilhança de β_1. O Score Test não necessita desse cálculo, sendo esse o fato de maior importância do estimador. Ele é dado por:

$$ST = \frac{\sum_{i=1}^{n} x_i * (y_i - \overline{y})}{\sqrt{\overline{y} * (1 - \overline{y}) * \sum_{i=1}^{n} (x_i - \overline{x})^2}} \quad (10)$$

Apesar de ST e W existirem, Hauck e Domer (1977) e Jennings (1986) estudaram a performance destes testes e verificaram que em certos casos eles falham e ambos autores indicam que o teste da razão de verossimilhança é o mais adequado.

3.1.4 Intervalos de Confiança

Em determinados casos é de interesse formular intervalos de confiança para $\widehat{\boldsymbol{\beta}}$. A base para construção deles é a mesma dos testes de significância, em particular o teste de Wald. O intervalo de 100(1-α)% de confiança para β_1 e β_0 são $\widehat{\beta_1} \pm z_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\beta_1})$ e $\widehat{\beta_0} \pm z_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\beta_0})$.

3.2 Regressão Logística Múltipla

Como visto até agora, foi introduzida a regressão logística no caso univariado. Porém, a força de uma técnica de modelagem consiste em modelar quantas variáveis forem necessárias, inclusive variáveis em diferentes escalas de mensuração. A abordagem de estimação e modelagem seguirá o mesmo procedimento usado na regressão logística simples.

3.2.1 Modelo

Considere o conjunto de p variáveis independentes descrita pelo vetor $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ e a probabilidade condicional de que a variável resposta está presente por $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$. O logito do modelo de regressão logística é dado pela equação

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_p * x_p \quad (11),$$

tal que o modelo de regressão logística fica

$$\pi(\boldsymbol{x}) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (12) \ .$$

No caso da inclusão de variáveis discretas de escala nominal é inapropriado usá-las como se fossem variáveis de escala intervalar. Os números usados para representa-las não possuem nenhuma significância numérica, eles são apenas identificadores. A maioria dos softwares estatísticos geram as variáveis identificadoras quando indicadas as variáveis com escala nominal. Em geral, se a variável de escala nominal possui k categorias, será necessário o uso de k-1 variáveis indicadoras para a variável em estudo.

3.2.2 Estimação dos Parâmetros

O método usado para estimação dos parâmetros será o mesmo do caso univariado, o método da máxima verossimilhança. A função de verossimilhança é a mesma da regressão logística simples com o fato de que $\pi(\boldsymbol{x})$ é definido como $\pi(\boldsymbol{x}) = \frac{e^{g(x)}}{1+e^{g(x)}}$.

Quando foi tratada regressão logística com uma variável independente, a abordagem do erro padrão dos estimadores não foi ampla. Como agora, o estudo foi generalizado para o caso multivariado, olha-se para este caso com maiores detalhes.

Como se pode observar em Rao (1973), o método de estimação das variâncias e covariâncias dos coeficientes estimados vem de uma teoria amplamente usada de estimação

por máxima verossimilhança. Essa teoria propõe que os estimadores são obtidos da matriz de segundas derivadas parciais da função de log verossimilhança e são da forma

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 * \pi_i * (1 - \pi_i) \quad (13)$$

e

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} * x_{il} * \pi_i * (1 - \pi_i) \quad (14)$$

para j, l = 0, 1, 2, ..., p onde π_i denota $\pi(\boldsymbol{x})$. A matriz (p+1)x(p+1) que contém o negativo dos termos das equações acima será denotada por $\boldsymbol{I}(\boldsymbol{\beta})$, que é a matriz de informação observada. As variâncias e covariâncias são obtidas da inversa da matriz que é definida como $Var(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$. Exceto em certas situações especiais, não é possível escrever uma expressão explícita dos elementos nessa matriz. Portanto, a notação Var(β_j) será usada para denotar o j-ésimo elemento da diagonal dessa matriz, que é a variância de $\widehat{\beta_J}$, e Cov(β_j, β_l) para denotar um elemento arbitrário fora da diagonal, que é a covariância de $\widehat{\beta_J}$ e $\widehat{\beta_l}$. Os estimadores de variância e covariância serão obtidos avaliando Var(β_j) em $\widehat{\boldsymbol{\beta}}$.

Uma formulação da matriz de informação que será útil na discussão de modelagem e avaliação da modelagem é $\hat{\boldsymbol{I}}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{X'VX}$ onde **X** é uma matriz n por (p+1) contendo os dados de cada variável explicativa e **V** uma matriz n por n com diagonal $\widehat{\pi_\iota} * (1 - \widehat{\pi_\iota})$. Ou seja,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad (15)$$

e a matriz **V** é

$$\boldsymbol{V} = \begin{bmatrix} \widehat{\pi_1} * (1 - \widehat{\pi_1}) & 0 & \cdots & 0 \\ 0 & \widehat{\pi_2} * (1 - \widehat{\pi_2}) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \widehat{\pi_n} * (1 - \widehat{\pi_n}) \end{bmatrix} \quad (16)$$

3.2.3Teste de Significância do Estimador

Como no caso univariado, a avaliação das variáveis explicativas que comporão o modelo segue da mesma forma. O teste da razão de verossimilhança é usado para avaliar os p coeficientes das variáveis explicativas e o teste é baseado na estatística G já abordada. A única diferença vem do fato de que os valores ajustados $\hat{\pi}$ são baseados nos (p+1) parâmetros, $\widehat{\boldsymbol{\beta}}$. Sob a hipótese nula de que os p coeficientes das covariáveis no modelo são nulos a distribuição de G será um qui-quadrado com p graus de liberdade.

A mesma abordagem é equivalente para o teste de Wald. Sob a hipótese de que um coeficiente em individual é igual à zero, a estatística $W = \frac{\widehat{\beta_J}}{\widehat{SE}(\widehat{\beta_J})}$ segue uma normal padrão.

Quando avaliadas as variáveis, sempre que uma variável do caso já discutido de escala nominal é incluída no modelo, todas as variáveis indicadoras dessa variável qualitativa devem compor o modelo. Isso levou a um problema devido ao fato do teste de Wald fornecer estimativas individuais para os coeficientes, porém a variável está decomposta em outras indicadoras, que neste teste são tratadas como variáveis diferentes. Então um teste análogo ao de Wald, porém multivariado, é definido por

$$W = \widehat{\boldsymbol{\beta}}'\left[\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'(\boldsymbol{X'VX})\widehat{\boldsymbol{\beta}} \quad (17)$$

que, sob a hipótese nula de que cada um dos p+1 coeficientes são iguais à zero, segue uma qui-quadrado com p+1 graus de liberdade.

3.2.4 Intervalos de Confiança

Para calcular os intervalos de confiança para cada coeficiente estimado, a abordagem usada no caso univariado será a mesma no caso multivariado. O intervalo de 100(1-α)% de confiança para β_i e β_0 são $\widehat{\beta_\iota} \pm z_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\beta_\iota})$ e $\widehat{\beta_0} \pm z_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\beta_0})$, para i=1,2, ..., p.

O intervalo de confiança para o logito do modelo é um pouco mais complicado devido ao fato de mais termos estarem envolvido no seu cálculo. Um meio de expressar o estimador logito é $\hat{g}(\boldsymbol{x}) = \boldsymbol{x'}\widehat{\boldsymbol{\beta}}$, onde $\widehat{\boldsymbol{\beta}}' = (\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \dots, \widehat{\boldsymbol{\beta}}_p)$ é o vetor dos p+1 coeficientes e o vetor $\boldsymbol{x'} = (x_0, x_1, x_2, \dots, x_p)$ representam a constante e os valores das p-covariáveis do modelo, onde x_o=1.

Sabendo que $\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X'VX})^{-1}$ segue que

$$\widehat{Var}[\hat{g}(\boldsymbol{x})] = \boldsymbol{x}'\widehat{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x} \quad (18)$$

Felizmente, os bons pacotes estatísticos com regressão logística possuem a opção de o usuário criar uma nova variável contendo os valores estimados da equação acima ou o erro padrão das covariáveis do banco de dados. Isso elimina a dificuldade computacional de calcular a equação acima e possibilita ao usuário de calcular rotineiramente os valores ajustados e o intervalo de confiança dos estimadores.

3.3 Interpretação do modelo de regressão logística ajustado

Para o estudo nesta seção, parte-se do pressuposto que um modelo de regressão logística foi ajustado e que todas as variáveis presentes no modelo são significantes clinicamente ou estatisticamente e que o modelo é adequado a partir de alguma medida estatística já vista. Basicamente a interpretação envolve duas questões: determinar a associação funcional entre a variável resposta e a variável independente, e definir apropriadamente a unidade de mudança para a variável independente.

Em regressão logística, o coeficiente angular representa a mudança no logito correspondente à mudança de uma unidade na variável independente, ou seja, $\beta_1 = g(x+1) - g(x)$. Interpretação adequada desse coeficiente em regressão logística depende de ser capaz de colocar significado na diferença indicada acima entre dois logitos. Para tal, cada caso será estudado adiante.

3.3.1 Variável explicativa dicotômica

Neste caso, considera-se que a variável explicativa é de escala nominal e dicotômica. Este caso será estudado primeiramente, pois fornece a fundamentação conceitual para as outras. Para prosseguir no resto do trabalho, sempre que uma variável for dicotômica ela será codificada em 0 e 1. Mais a frente verifica-se a importância de ressaltar que esta codificação será utilizada.

A diferença do logito de uma variável para $x = 0$ e $x = 1$ é dado por

$$g(1) - g(0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1 \quad (19).$$

Essa equação é usada para enfatizar que o primeiro passo para se interpretar o efeito da covariável é expressar a diferença do logito em termos do modelo, que neste caso é igual a β_1.

Para interpretar este resultado faz-se necessária a discussão de uma medida de associoção, a razão de chances.

A chance de uma variável resposta pode ser definida como $\pi(x)/(1-\pi(x))$, tal que se $x = 1$ a variável resposta está presente e se $x = 0$ a variável resposta não está presente. A razão de chances (OR) é definida pela razão da chance para $x = 1$ por $x = 0$, e é dada pela equação

$$OR = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} \quad (20).$$

E substituindo as expressões do modelo de regressão logística obtém-se

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1} \quad (21)\,.$$

Então, para uma regressão logística com uma variável explicativa dicotômica codificada como 0 e 1 a relação entre a razão de chances e o coeficiente de regressão é

$$OR = e^{\beta_1} \quad (22).$$

Essa simples relação entre o coeficiente e a razão de chances é a razão fundamental do porquê que a regressão logística mostra-se uma poderosa ferramenta de pesquisa analítica.

A razão de chances é uma medida de associação que obteve grande uso, especialmente na área de saúde, porque ela aproxima o quão mais provável (ou improvável) é que a variável resposta esteja presente naqueles que $x = 1$ do que aqueles que $x = 0$. Sua interpretação é baseada no fato de que, em vários casos, ela aproxima o risco relativo. Esse parâmetro é definido pela razão $\pi(1)/\pi(0)$. Segue da equação da razão de chances que ela aproxima o risco relativo se $[1-\pi(0)]/[1-\pi(1)] \approx 1$. Isso se mantém se $\pi(x)$ é pequeno para $x = 0$ e $x = 1$.

Usualmente, a razão de chances é o parâmetro de interesse em uma regressão logística devido à sua fácil interpretação. Porém, a sua estimação, $\widehat{OR}$, tende a ter uma distribuição viesada. O viés da distribuição amostral de $\widehat{OR}$ é devido ao fato de que os possíveis valores variam de 0 a ∞. Em teoria, para grandes amostras, $\widehat{OR}$ segue uma distribuição normal. Mas,

essa necessidade de uma grande amostra geralmente não é satisfeita na maioria dos estudos. Portanto, as inferências são baseadas na distribuição amostral de $\ln \widehat{OR} = \widehat{\beta_1}$, que segue uma distribuição normal para valores amostrais bem menores.

Então, um intervalo de 100x(1-α)% de confiança para β_1 é definido por

$$\exp\left[\widehat{\beta_1} \pm z_{1-\frac{\alpha}{2}} * \widehat{SE}\left(\widehat{\beta_1}\right)\right] \quad (23).$$

Resumindo, como discutido anteriormente, quando a variável explicativa é dicotômica o parâmetro de interesse no estudo é a razão de chances, que pode ser obtido pela estimação do coeficiente da regressão logística independente de como foi codificada a variável. Essa relação entre a razão de chances e a regressão logística que servirá de base para a continuidade do estudo.

3.3.2 Variável Explicativa Politômica

Em alguns casos a medida da variável nominal não é em apenas duas categorias e sim algum valor k > 2 de categorias para a variável. Para isso, faz-se uso do mesmo método utilizado anteriormente com a variável explicativa binária, criam-se variáveis identificadoras.

Para cada variável com k > 2 categorias, criam-se k-1 variáveis identificadoras. Determina-se qual categoria será a de referência e para ela todas as k-1 variáveis criadas são iguais a zero, para a próxima categoria uma das variáveis é 1 e as outras 0 e assim até todas as k categorias estarem bem definidas pelas novas variáveis indicadoras.

Para obter intervalos de confiança para as estimativas da razão de chances (o coeficiente da regressão logística) a mesma abordagem usada em variáveis binárias se aplica. Ou seja, os limites para um intervalo de confiança de 100x(1-α)% para β_j é dado por

$$\exp\left[\widehat{\beta_J} \pm z_{1-\frac{\alpha}{2}} * \widehat{SE}\left(\widehat{\beta_J}\right)\right] \quad (24).$$

3.3.3 Variável Explicativa Contínua

Para este tipo de interpretação, assume-se que o logito é linear na variável, ou seja, $g(x) = \beta_0 + \beta_1 * x$. Segue que o coeficiente angular, β_1, da a mudança na log chance para cada aumento em uma unidade em *x*, isto é, $\beta_1 = g(x+1) - g(x)$, para qualquer valor de x.

Porém, em muitos casos saber essa mudança em apenas 1 unidade em *x* não é interessante, as variáveis contínuas podem ter diferentes alcances e diferentes interpretações na unidade de variação. Por isso a necessidade de quando a variável explicativa for contínua a análise ser feita com mudança de *c* unidades em *x*.

O log da razão de chances para uma variação de *c* unidades em *x* é obtido pela diferença dos logitos $g(x+c) - g(x) = c\beta_1$ e a razão de chances retirando a exponencial do resultado da diferença, $OR(c) = OR(x+c, x) = \exp(c\beta_1)$. Para obter intervalos de confiança basta substituir β_1 pela sua estimativa de máxima verossimilhança e considerar o peso *c* no intervalo. Com isso temos que o intervalo de confiança de 100x(1-α)% para OR(c) é

$$\exp\left[\widehat{c\beta_1} \pm z_{1-\frac{\alpha}{2}} * c * \widehat{SE}\left(\widehat{\beta_1}\right)\right] \quad (25)\,.$$

Como o valor de *c* é arbitrário, a análise sempre deverá explicitar qual o valor de *c* utilizado e o por quê dele ser interessante para a análise.

3.3.4 O Modelo Multivariado

Até agora a interpretação baseou-se em uma série de modelos univariados, porém isso raramente produz interpretações corretas se imaginar que muitas das variáveis explicativas possuem associações com outras e podem ter diferentes distribuições dentro de cada nível da variável resposta. O objetivo dessa análise é ajustar estatisticamente os efeitos estimados de cada variável no modelo para diferenças nas associações e distribuições entre as outras covariáveis.

Para explicitar melhor a interpretação neste caso, faz-se necessário uma análise de interação entre as variáveis e variáveis confundidoras.

3.3.5 Interação e Variáveis Confundidoras

O termo confundidor é utilizado para descrever uma covariável que é associada tanto com a variável resposta quanto com uma variável independente primária ou que seja fator de risco. Quando as duas associações estão presentes, então a interação do fator de risco com a variável resposta é dita confusa. Quando não há interação, a associação da covariável com a variável resposta é a mesma para cada nível do fator de risco, a abordagem é a mesma dita anteriormente para variáveis independentes dicotômicas, politômicas e contínuas, basta obter

os valores ajustados da razão de chances que diferem apenas na característica de interesse e manter todas as outras variáveis constantes.

Quando existe interação, a associação entre o fator de risco e a variável resposta difere ou depende de cada nível da covariável. O modelo mais simples e comumente usado que inclui interações é um no qual considera que o logito também é linear na variável confundidora para o outro grupo, mas com um coeficiente angular diferente. Ou seja, dois logitos (no caso de dois grupos) lineares com mesmo intercepto e inclinações diferentes. Para ajudar no entendimento, o gráfico abaixo mostra a situação descrita.

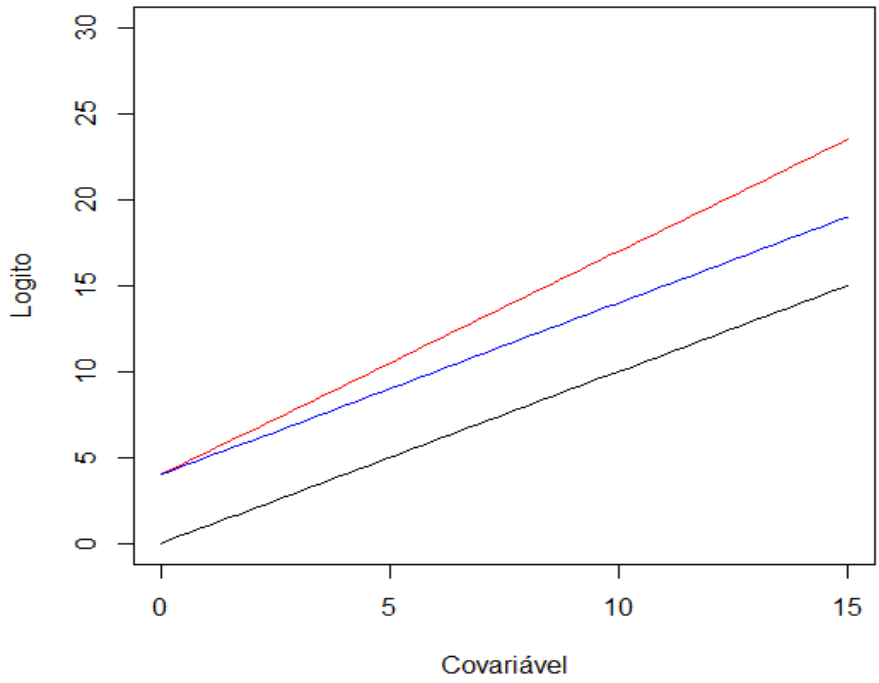

Figura 1 – Gráfico dos logitos de três diferentes modelos mostrando a presença e ausência de interação.

A Figura 1 mostra três diferentes logitos hipotéticos. Suponha que neste caso para melhor exemplificar o fator de risco possua apenas duas categorias e que l_1 corresponda ao logito de um dos grupos do fator de risco em função de uma covariável, indicado pela cor preta, e l_2 ao logito do outro grupo, indicado pela cor azul. Como as linhas são paralelas, isso indica que a interação entre o fator de risco e a variável resposta é a mesma independente da covariável. Neste caso, não existe interação e o logaritmo da razão de chances para o fator de risco, controlando a covariável, é dado pela diferença $l_2 - l_1$, que é igual a distância vertical entre as duas linhas, constante para toda a covariável.

Agora suponha que ao invés de l_1 e l_2 serem os logitos do fator de risco em função da covariável, os logitos sejam dados por l_2 (azul) e l_3, indicado pela cor vermelha. Note que os logitos possuem inclinações diferentes e, quando isso acontece, indica que o fator de risco está associado à covariável. A estimativa do log da razão de chances é, também, indicada pela distância vertical dos logitos, $l_2 - l_3$, mas agora depende de qual nível da covariável se trata. Portanto, não deve-se estimar a razão de chances antes de determinar em qual nível da covariável está sendo feita a análise. Quando um modelo é ajustado, a interação é adicionada criando uma variável que é igual ao produto do valor do fator de risco e do valor da covariável.

Portanto, percebe-se que determinar se uma covariável é associada com o fator de risco (modificadora de efeito) e/ou confundidora envolve diferentes questões. O gráfico da Figura 1 mostra que determinar modificação de efeito envolve diretamente a estrutura paramétrica do logito, enquanto verificar se a variável independente é confundidora envolve dois aspectos. Primeiro, a covariável tem que ser associada com a variável resposta, isso implica que o logito deve ter um coeficiente angular diferente de zero. Segundo, a covariável tem que ser associada com o fator de risco.

Na prática, para checar se a covariável é confundidora é comparar os coeficientes estimados para a variável do fator de risco de modelos que contém e não contém a covariável. Qualquer mudança importante no coeficiente estimado para o fator de risco sugere que a covariável é confundidora e deve ser incluída no modelo independentemente de sua significância estatística. Por outro lado, só define-se a variável como modificadora de efeito quando o termo adicionado ao modelo da interação é significante tanto estatisticamente quanto clinicamente. Assim que determina-se uma variável como modificadora de efeito, o *status* de confundidora torna-se obsoleto já que a estimação do efeito do fator de risco depende de valores específicos da covariável.

3.3.6 Estimação da Razão de Chances na Presença de Interação

Na seção anterior, mostra-se que na presença de interação entre um fator de risco e outra variável, a estimativa da razão de chances para o fator de risco depende do valor da variável que está interagindo com ele. Nestes casos a estimação da razão de chances pode não estar correta apenas aplicando a exponencial no coeficiente estimado. Um método que levará sempre ao estimador correto baseado no seu modelo possui três passos. Primeiro deve-se

escrever as expressões do logito nos dois níveis que serão comparados do fator de risco; segundo é simplificar algebricamente a diferença dos logitos e computar o seu valor; e, por fim, aplicar a exponencial no valor encontrado no segundo passo.

Para facilitar a abordagem de como estimar a razão de chances e construir intervalos de confiança aplica-se o método acima apenas com duas variáveis e sua interação. O fator de risco será denotado como F e a variável X. Para avaliações com mais variáveis e interações basta extrapolar a metodologia a seguir. O logito para o modelo avaliado em $F = f$ e $X = x$ é

$$g(f, x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f \times x \qquad (26).$$

O objetivo é a razão de chances comparando dois níveis de F, $F = f_1$ e $F = f_0$, onde $X = x$. Seguindo o procedimento de três passos têm-se

$$g(f_1, x) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x$$

e

$$g(f_0, x) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x$$

Depois computa-se e simplifica-se a diferença para obter o log da razão de chances

$$\begin{aligned} \ln[OR(F = f_1, F = f_0, X = x)] &= g(f_1, x) - g(f_0, x) \\ &= (\beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x) - (\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x) \\ &= \beta_1 (f_1 - f_0) + \beta_3 x (f_1 - f_0) \qquad (27). \end{aligned}$$

Por último aplica-se a exponencial no valor encontrado em (27)

$$OR = \exp[\beta_1 (f_1 - f_0) + \beta_3 x (f_1 - f_0)] \qquad (28).$$

Interessante notar que a expressão para o log da razão de chances (27) não se simplifica a apenas um coeficiente, ela envolve dois coeficientes, a diferença nos valores do fator de risco e a variável de interação. Obviamente, o estimador da razão de chances é obtido substituindo os parâmetros pelo seus estimadores.

Para obter os limites do intervalo de confiança para o estimador da razão de chances encontrado, a abordagem é a mesma para modelos sem interação. E, para isso, deve-se

estimar a variância do estimador do log da razão de chances em (27). Usando métodos para calcular a variância de uma soma, o seguinte estimador é obtido

$$\widehat{Var}\{\ln[\widehat{OR}(F=f_1,F=f_0,X=x)]\} = (f_1-f_0)^2 \times \widehat{Var}(\hat{\beta}_1) + [x(f_1-f_0)]^2 \times \widehat{Var}(\hat{\beta}_3) + 2x(f_1-f_0)^2 \times \widehat{Cov}(\hat{\beta}_1,\hat{\beta}_3) \qquad (29).$$

Substituindo os estimadores da variância e covariância em (29), obtem-se o estimador da variância do log da razão de chances. O intervalo de $100 \times (1-\alpha)\%$ de confiança para o log da razão de chances é

$$[\hat{\beta}_1(f_1-f_0)+\hat{\beta}_3 x(f_1-f_0)] \pm z_{1-\frac{\alpha}{2}}\widehat{SE}\{\ln[\widehat{OR}(F=f_1,F=f_0,X=x)]\} \qquad (30).$$

Para obter o intervalo de confiança da razão de chances basta aplicar a exponencial em (30). Note que no caso do fator de risco ser binário, $(f_1-f_0)=1$, o que simplifica bastante as expressões (29) e (30).

3.4 Ajustando Modelos de Regressão Logística para Dados de Amostras Complexas

Como anunciado na introdução deste trabalho, em muitas ocasiões por motivos de custo ou de facilidade, os dados que pesquisadores obtêm não são provenientes de amostras aleatórias simples, eles vêm de um plano amostral complexo, como estratificação e conglomeração. Por isso o foco de abordar a análise da regressão logística nestes casos.

Como Roberts, Rao e Kumar (1987) discutem, a ideia principal é definir uma função que aproxima a função de verossimilhança da população finita amostrada com uma função de verossimilhança formada pela amostra observada e os pesos amostrais conhecidos. Suponha que a população possa ser dividida em k = 1, 2, ..., K estratos, j = 1, 2, ..., M_k unidades amostrais primárias em cada estrato e i = 1, 2, ..., N_{kj} elementos na kj-ésima unidade primária amostral. Suponha também que os dados observados consistem de n_{kj} elementos das m_k unidades primárias amostrais do estrato k. O número total de observações é dado por $n = \sum_{k=1}^{K}\sum_{j=1}^{m_k} n_{kj}$, os pesos amostrais conhecidos da kji-ésima observação por w_{kji}, o vetor de covariáveis $\boldsymbol{x_{kji}}$ e a variável resposta binária por y_{kji}. A função de log-verossimilhança aproximada é

$$\sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}}\left[w_{kji} * y_{kji}\right] * \ln\left[\pi\left(\boldsymbol{x}_{kji}\right)\right] + \left[w_{kji} * \left(1 - y_{kji}\right)\right] * \ln\left[1 - \pi\left(\boldsymbol{x}_{kji}\right)\right] \quad (31)$$

derivando em respeito aos coeficientes desconhecidos da regressão tem-se o vetor de *p + 1* equações

$$\boldsymbol{X'W}(\boldsymbol{y} - \boldsymbol{\pi}) = \mathbf{0} \quad (32),$$

Onde $\boldsymbol{X}$ é a *n* x *(p + 1)* matriz de valores das covariáveis, $\boldsymbol{W}$ é a *n* x *n* matriz diagonal contendo os pesos, **y** é o *n* x *1* vetor das observações da variável resposta e $\boldsymbol{\pi} = \left(\pi(\boldsymbol{x}_{111}), \ldots, \pi\left(\boldsymbol{x}_{Km_kn_{kj}}\right)\right)$ é o *n* x *1* vetor das probabilidades logísticas.

O problema aparece na hora de obter o estimador correto da matriz de covariâncias do estimador dos coeficientes. Uso errôneo de softwares estatísticos com matriz de pesos $\boldsymbol{W}$ levariam a estimações na matriz $(\boldsymbol{X'DX})^{-1}$ onde $\boldsymbol{D} = \boldsymbol{WV}$ é uma matriz diagonal *n* x *n* com elemento geral $w_{kji} * \hat{\pi}\left(\boldsymbol{x}_{kji}\right)\left[1 - \hat{\pi}\left(\boldsymbol{x}_{kji}\right)\right]$. O estimador correto é

$$\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X'DX})^{-1}\boldsymbol{S}(\boldsymbol{X'DX})^{-1} \quad (33)\,,$$

onde S é o estimador agrupado intra-estrato da matriz de covariâncias do lado esquerdo da equação (32). Denote um elemento geral no vetor em (32) como $\boldsymbol{z}'_{kji} = \boldsymbol{x}'_{kji}w_{kji}\left(y_{kji} - \pi\left(\boldsymbol{x}_{kji}\right)\right)$, o somatório para todas as n_{kj} unidades amostradas na j-ésima unidade primária amostral do k-ésimo estrato como $\boldsymbol{z}_{kj} = \sum_{i=1}^{n_{kj}}\boldsymbol{z}_{kji}$ e sua média específica do estrato como $\bar{\boldsymbol{z}}_k = \frac{1}{m_k}\sum_{j=1}^{m_k}\boldsymbol{z}_{kj}$. O estimador intra-estrato para o k-ésimo estrato é

$$\boldsymbol{S}_k = \frac{m_k}{m_k - 1}\sum_{j=1}^{m_k}(\boldsymbol{z}_{kj} - \bar{\boldsymbol{z}}_k)(\boldsymbol{z}_{kj} - \bar{\boldsymbol{z}}_k)' \quad (34).$$

O estimador agrupado é $\boldsymbol{S} = \sum_{k=1}^{K}(1 - f_k)\boldsymbol{S}_k$. A quantidade $\left(1 - f_k\right)$ é chamada de fator de correção para população finita onde $f_k = \frac{m_k}{M_k}$ é a razão do número de unidades amostrais primárias observadas pelo número total de unidades amostrais primárias no estrato *k*. Em alguns casos não pode-se determinar qual valor de M_k, então é comum assumir que ele é grande suficiente para que o fator de correção para população finita seja igual a um.

A função de verossimilhança (31) é apenas uma aproximação. Mesmo assim, espera-se que as inferências sejam baseadas nas estatísticas de Wald como foi a abordagem até agora. Porém, Korn e Graubard (1990) mostraram que quando os dados são provenientes de delineamentos complexos de populações finitas, o uso de um teste de Wald modificado juntamente com a distribuição F levam a testes com maior aderência com o nível alfa estabelecido.

Seja *W* a estatística de Wald para testar que todos os p coeficientes angulares do modelo ajustado são iguais a zero, segue que

$$W = \widehat{\boldsymbol{\beta}}' \left[\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)_{pxp}\right]^{-1} \widehat{\boldsymbol{\beta}} \quad (35),$$

Onde $\widehat{\boldsymbol{\beta}}$ denota o vetor dos *p* coeficientes angulares e $\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)_{pxp}$ a sub-matriz *p* x *p* obtida da matriz completa *(p +1)* x *(p +1)* da equação (28). O p-valor é encontrado usando uma distribuição qui-quadrado com *p* graus de liberdade como $p - valor = P[\chi^2(p) \geq W]$.

A estatística de Wald modificada é

$$F = \frac{s - p + 1}{sp} W \quad (36),$$

onde $s = (\sum_{k=1}^{K} m_k) - K$ é o número total de unidades amostrais primárias amostradas menos o número de estratos. O p-valor é encontrado usando uma distribuição F com *p* e *(s – p + 1)* graus de liberdade como $p - valor = P[F(p, s - p + 1) \geq F]$.

4 Banco de Dados

4.1 Introdução

O estudo será realizado em uma pesquisa de doutorado feita por Taís Galvão, aluna do programa de doutorado em Ciências da Saúde, com o objetivo de estimar a prevalência e os fatores associados à depressão autorreferida em adultos residentes em Brasília. A pesquisa abordou adultos entre 18 e 65 anos moradores em Brasília no segundo semestre de 2012. As regiões administrativas de Brasília que foram consideradas para a amostra foram: Asa Norte, Asa Sul, Brazlândia, Candangolândia, Ceilândia, Cruzeiro, Gama, Guará, Lago Norte, Lago Sul, Núcleo Bandeirante, Paranoá, Planaltina, Recanto das Emas, Riacho Fundo, Samambaia, Santa Maria, São Sebastião, Sobradinho e Taguatinga. O delineamento amostral foi feito para que todas as classes sociais tivessem representatividade na amostra.

4.2 Seleção da Amostra

Segundo o censo demográfico de 2010, Brasília tem 1.702.419 residentes entre 18 a 65 anos. A estimativa de depressão autorreferida usada para calcular a amostra foi de 10%, com nível de confiança de 95% e erro de 1,5%. Com estes dados obteve-se o tamanho de amostra igual a 1.536 pessoas. A este número foi adicionado 20% a mais caso seja necessária alguma compensação. Com isso a amostra total é de 1.843 pessoas.

A amostragem realizada foi probabilística por conglomerados em dois estágios. Considerou-se apenas os 3.886 setores censitários com mais de 200 moradores de Brasília definido pelo IBGE (Instituto Brasileiro de Geografia e Estatística) e foram sorteados 182 setores primários e 38 setores de reposição. Para cada setor primário selecionado foram sorteados 10 domicílios e entrevistado um membro da família.

4.3 Coleta de Dados

O instrumento de coleta consistiu de um questionário semiestruturado e pré-codificado, composto por quatro domínios: (i) socioeconômico (sexo, idade, estado civil, número de residentes, escolaridade e ocupação); (ii) situação de saúde autorreferida (depressão, diabetes, hipertensão, doença cardiovascular, doença respiratória, outras doenças crônicas, acesso a serviços de saúde e avaliação subjetiva do estado de saúde); (iii) consumo de medicamentos (nome comercial, disponibilidade da embalagem, dose, tempo de consumo,

responsável pela indicação e forma de acesso); e (iv) informações para o critério de classificação econômica.

Entrevistadores com experiência em coleta de dados para pesquisas quantitativas preencheram o papel do questionário, no domicílio do entrevistado, após assinatura do Termo de Consentimento Livre e Esclarecido. A compreensão do instrumento foi avaliada por meio de pré-teste compreendendo 150 participantes e para garantir a fidedignidade dos dados coletados, 20% das entrevistas foram auditadas por meio de contato telefônico com o entrevistado.

A fim de minimizar erros de codificação, todos os dados tabulados foram conferidos por duas pessoas com o questionário original e foram excluídos aqueles que apresentavam erro de preenchimento, dados incompletos ou que sugeriram algum viés de memória.

5 Resultados

Para análise e resultados do banco de dados o software estatístico *Statistical Analysis System – SAS* foi utilizado e toda a programação estará disponível em anexo neste trabalho. Inicialmente será feita uma análise descritiva dos dados e, após esta análise, um modelo de regressão logística será ajustado.

As variáveis que serão consideradas na análise são: renda familiar, escolaridade, ocupação, gênero, faixa etária, estado conjugal, número de morado

res, diabetes, hipertensão, depressão, problema cardíaco, problemas respiratórios, outros problemas crônicos, plano privado de saúde, consulta médica, hospitalização, uso de antidepressivo, mobilidade, cuidado próprio, atividades cotidianas, dor e ansiedade/depressão. Totalizando um total de 22 variáveis consideradas no estudo.

5.1 Análise Descritiva

Começando a interpretação dos resultados pela análise descritivas dessas variáveis, o procedimento *surveyfreq* foi utilizado no SAS, pois pelos dados da amostra ele estima a frequência populacional e fornece intervalos de confiança para cada variável, além do gráfico que auxilia a compreensão. Para algumas variáveis mais relevantes alguns breves comentários serão feitos a respeito dos resultados.

Tabela 1: Frequências da variável classe social.

Classe social					
Classe social	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
D-E	166	9.4820	0.9725	7.5632	11.4009
C	868	47.5486	1.9694	43.6626	51.4346
B	624	34.4294	1.6722	31.1299	37.7289
A2	134	7.1448	1.0454	5.0821	9.2075
A1	28	1.3952	0.4630	0.4815	2.3088
Total	1820	100.000			

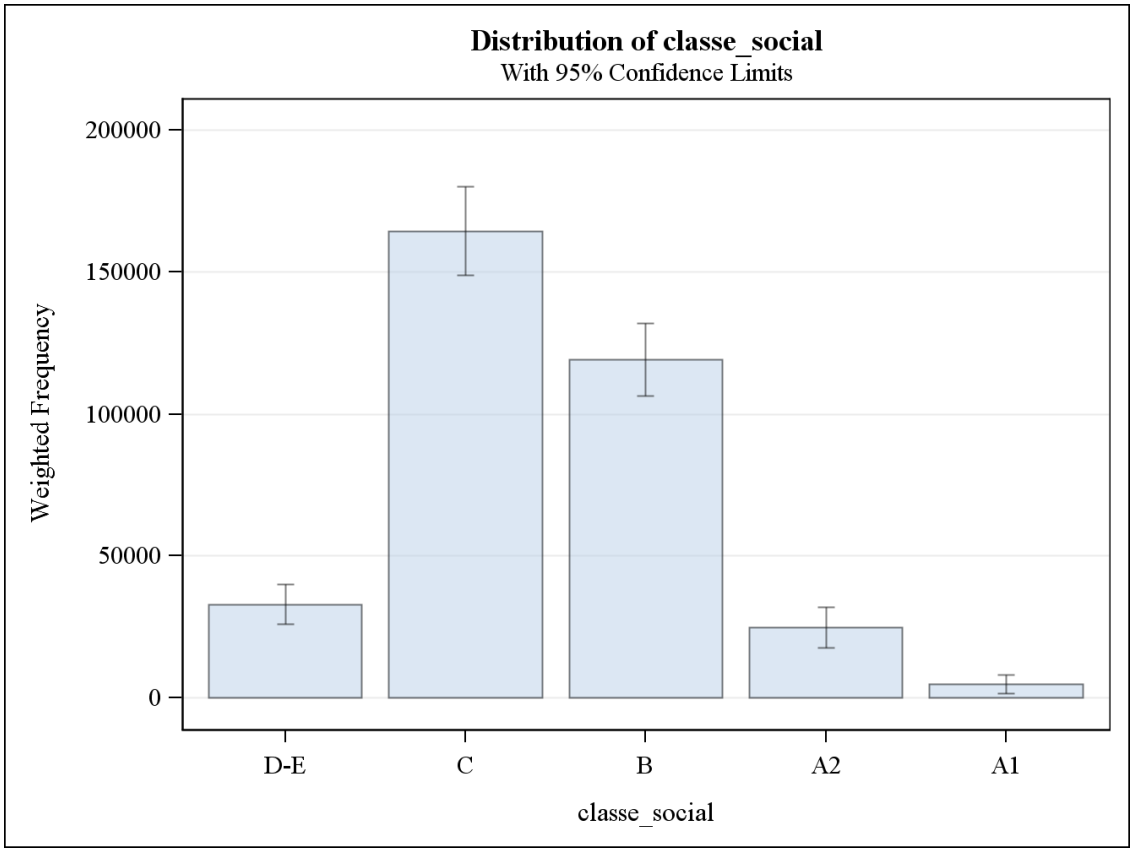

Figura 2 – Gráfico das frequências ponderadas da variável classe social.

Observando os dados da amostra sobre a variável classe social, observa-se como é a pirâmide social da população do DF. A maioria encontra-se nas classes C e B e percebem-se poucas pessoas que ganham acima de R$ 14.000,00 por mês.

Tabela 2: Frequências da variável escolaridade.

Escolaridade					
Escolaridade	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Baixa	483	26.5739	1.4413	23.7300	29.4179
Média Baixa	394	21.6419	1.2041	19.2660	24.0177
Média Alta	627	34.4259	1.3405	31.7809	37.0709
Alta	316	17.3583	1.6245	14.1530	20.5636
Total	1820	100.000			

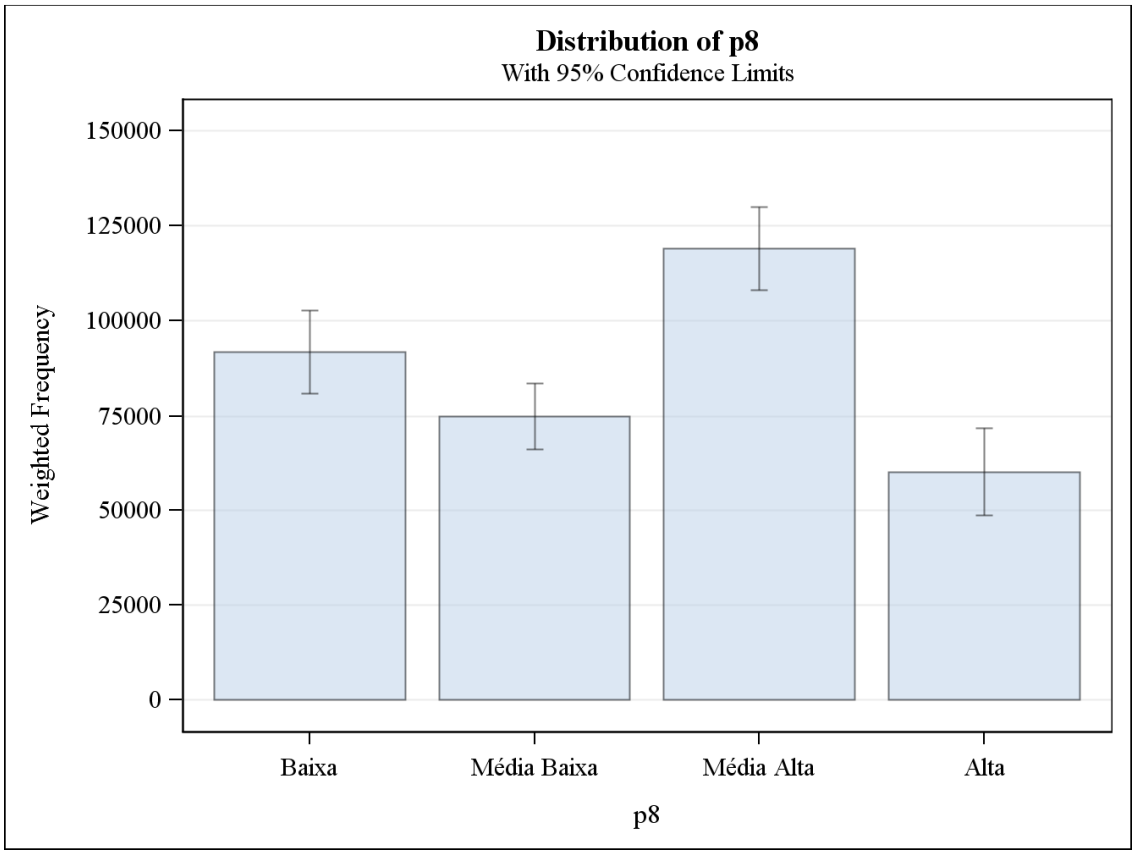

Figura 3 – Gráfico das frequências ponderadas da variável escolaridade.

A variável escolaridade foi classificada em quatro categorias: baixa (analfabeto + 1º grau incompleto), média baixa (1º grau completo + 2º grau incompleto), média alta (2º grau completo + 3º grau incompleto) e alta (3º grau completo + pós-graduação). Nela percebe-se uma maior simetria, com a escolaridade média alta sendo a mais representativa. Porém, é notável que ainda existem muitas pessoas que não possuem o ensino médio completo.

Tabela 3: Frequências da variável ocupação.

Ocupação					
Ocupação	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	1380	74.7397	1.5273	71.7261	77.7533
Não	440	25.2603	1.5273	22.2467	28.2739
Total	1820	100.000			

Considerou-se como uma pessoa não ocupada: desempregado, aposentado e não trabalha; como uma pessoa ocupada: servidor, trabalho informal, trabalho doméstico e autônomo. Pela amostra verifica-se que o percentual de pessoas ocupadas da população está entre 71% e 78% com 95% de confiança. Portanto, o percentual de não-ocupados ainda é alto.

Tabela 4: Frequências da variável gênero.

Gênero					
Gênero	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Masculino	731	40.6681	1.5616	37.5869	43.7493
Feminino	1089	59.3319	1.5616	56.2507	62.4131
Total	1820	100.000			

A amostra revela que a população do DF é predominantemente mulher (60% aproximadamente). Isso indica que a população do gênero masculino morreu mais, pois sabe-se que nascem um pouco mais de homens do que mulheres no Brasil (105 homens a cada 100 mulheres). Para que esse nível se mantenha igual alguma política pública para controle da mortalidade nos homens deve ser adotada.

Tabela 5: Frequências da variável faixa etária.

Faixa etária					
Faixa Etária	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
18 - 25	784	43.4585	1.3279	40.8384	46.0786
26 - 35	637	35.1258	1.3819	32.3992	37.8524
36 - 45	399	21.4157	1.1353	19.1755	23.6559
Total	1820	100.000			

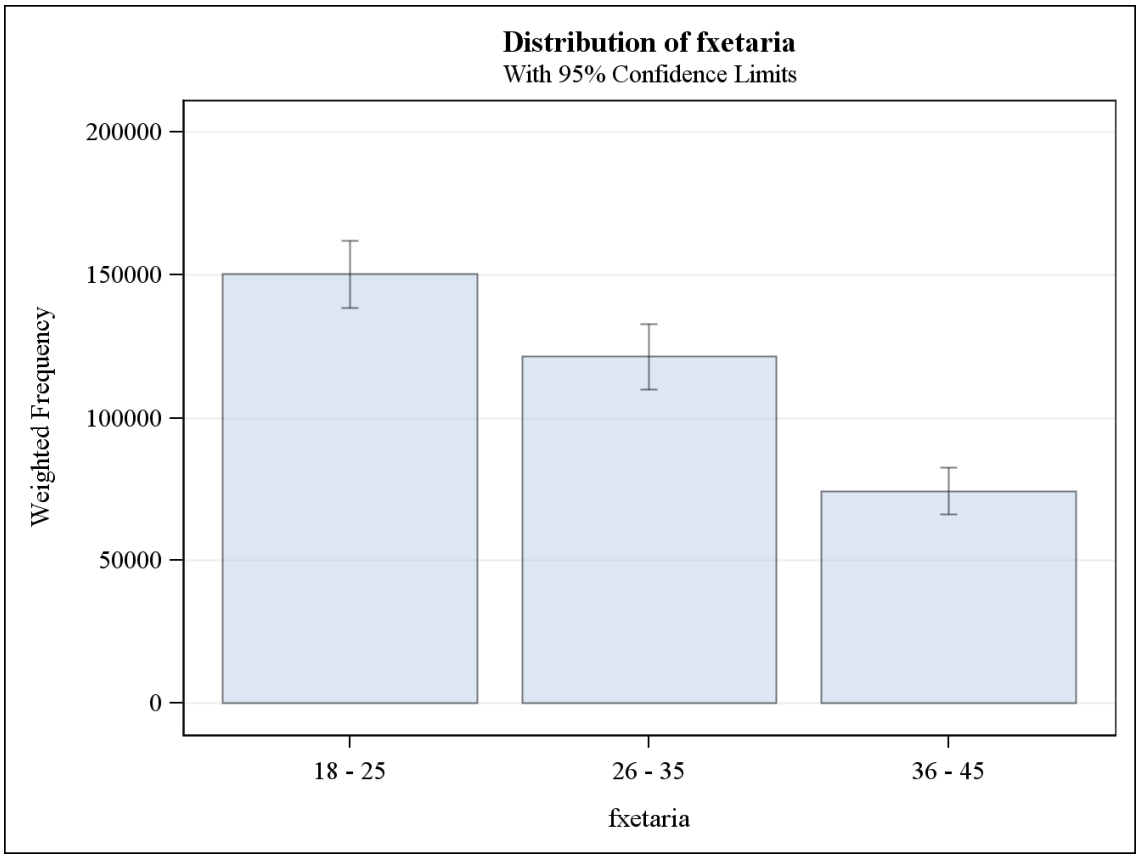

Figura 4 – Gráfico das frequências ponderadas da variável faixa etária.

Pelo gráfico a estrutura etária da população adulta do DF é analisada. Nela observa-se que trata de uma população jovem, com sua maioria entre 18 e 25 anos e a minoria entre 36-45.

Tabela 6: Frequências da variável estado conjugal.

Estado conjugal					
Estado conjugal	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Não-casado	868	47.8112	1.4300	44.9895	50.6329
Casado	952	52.1888	1.4300	49.3671	55.0105
Total	1820	100.000			

Para análise do estado conjugal, foram feitas duas categorias: não-casado e casado. Não-casado são as pessoas solteiras, separadas ou divorciadas ou viúvas, já os casados são as uniões consensuais e os casamentos. Vê-se que a divisão está bem equilibrada, 52% de casados contra 48% de não-casados.

Tabela 7: Frequências da variável número de moradores.

Número de moradores					
Número de moradores	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
1 a 3 pessoas	774	42.3366	1.3082	39.7554	44.9178
4 a 5 pessoas	792	44.2078	1.3191	41.6050	46.8107
6 em diante	254	13.4556	0.9371	11.6066	15.3046
Total	1820	100.000			

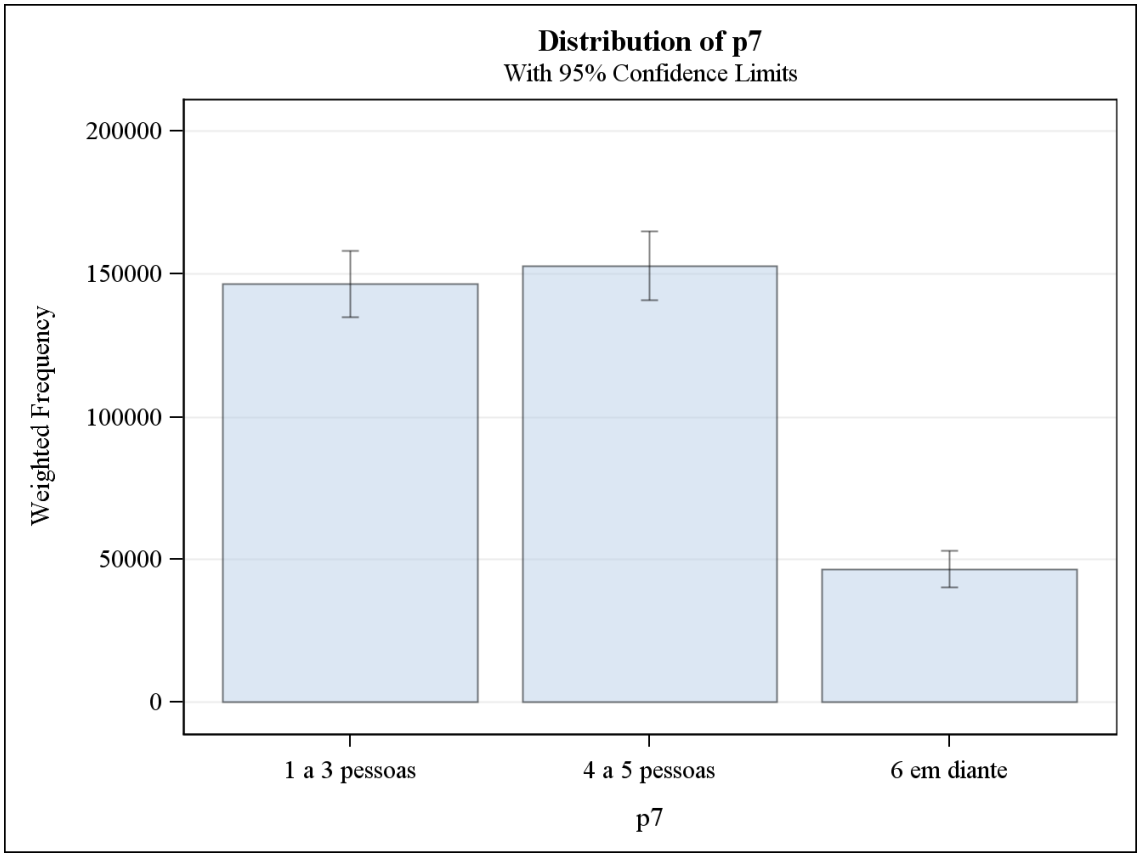

Figura 5 – Gráfico das frequências ponderadas da variável número de moradores.

A tabela e o gráfico de número de moradores no domicílio mostram que as famílias residentes são em sua maioria de pequenas para médias, com 5 pessoas no máximo.

A seguir, seguem seis tabelas com informações sobre as frequências e porcentagens com intervalos de confiança de doenças crônicas coletadas da amostra.

Tabela 8: Frequências da variável diabetes.

Diabetes					
Diabetes	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	177	10.0508	0.7900	8.4921	11.6096
Não	1541	89.9492	0.7900	88.3904	91.5079
Total	1718	100.000			
Frequência Faltante = 102					

Tabela 8: Frequências da variável hipertensão.

Hipertensão					
Hipertensão	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	383	21.5000	1.1336	19.2632	23.7368
Não	1379	78.5000	1.1336	76.2632	80.7368
Total	1762	100.000			
Frequência Faltante = 58					

Tabela 9: Frequências da variável depressão.

Depressão					
Depressão	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	218	12.7710	1.0043	10.7893	14.7527
Não	1541	87.2290	1.0043	85.2473	89.2107
Total	1759	100.000			
Frequência Faltante = 61					

Tabela 10: Frequências da variável problema cardíaco.

Problema cardíaco					
Problema cardíaco	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	116	6.9280	0.6854	5.5757	8.2804
Não	1609	93.0720	0.6854	91.7196	94.4243
Total	1725	100.000			
Frequência Faltante = 95					

Tabela 11: Frequências da variável problema respiratório.

Problema respiratório					
Problema respiratório	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	133	7.2965	0.7287	5.8586	8.7343
Não	1634	92.7035	0.7287	91.2657	94.1414
Total	1767	100.000			
Frequência Faltante = 53					

Tabela 12: Frequências da variável outros problemas crônicos.

Outros problemas crônicos					
Outros problemas	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	146	8.0032	0.7327	6.5575	9.4490
Não	1674	91.9968	0.7327	90.5510	93.4425
Total	1820	100.000			

Para as seis variáveis em análise temos que 10% possui diabetes, 22% hipertensão, 13% depressão, 7% problemas cardíacos, 7% problemas respiratórios e 8% outros problemas crônicos. Como tratam-se de doenças altamente perigosas ao ser humano, as que obtiveram percentual maior que 10% podem ser preocupantes por não se tratarem mais de casos raros, principalmente a hipertensão por já atingir 22% da população adulta.

Tabela 13: Frequências da variável plano privado de saúde.

Plano privado de saúde					
Plano privado	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	503	27.7074	1.9377	23.8839	31.5308
Não	1317	72.2926	1.9377	68.4692	76.1161
Total	1820	100.000			

Pela tabela acima chama-se atenção que apenas 28% da população adulta do DF possuem plano privado de saúde, indicando que muitos ainda recorrem ao sistema público de saúde ou em consultas particulares.

Tabela 14: Frequências da variável consulta médica.

Consulta Médica					
Consulta médica	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	766	42.0789	1.5230	39.0738	45.0840
Não	1054	57.9211	1.5230	54.9160	60.9262
Total	1820	100.000			

Tabela 15: Frequências da variável hospitalização.

Hospitalização					
Hospitalização	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	181	9.7371	0.7928	8.1728	11.3014
Não	1639	90.2629	0.7928	88.6986	91.8272
Total	1820	100.000			

Tabela 16: Frequências da variável uso de antidepressivos.

Uso de antidepressivo					
Uso de antidepressivo	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Sim	32	1.6436	0.3369	0.9789	2.3084
Não	1788	98.3564	0.3369	97.6916	99.0211
Total	1820	100.000			

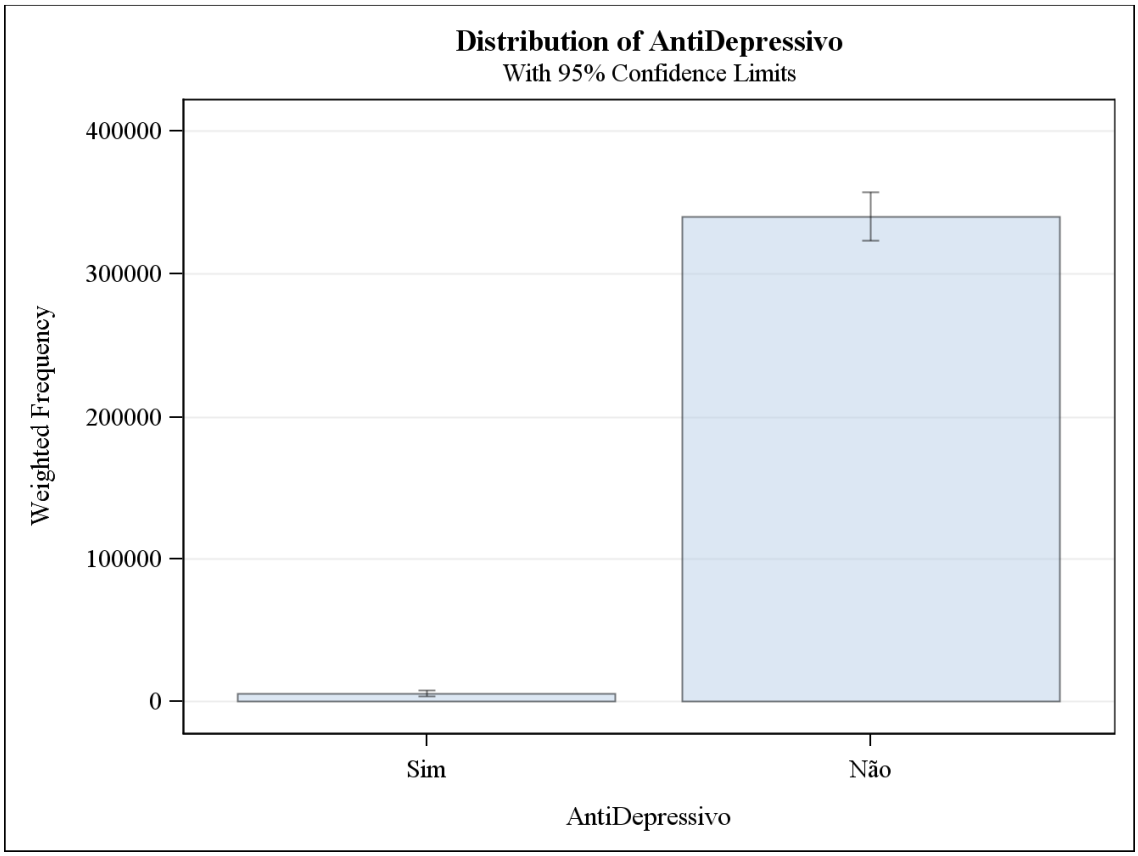

Figura 6 – Gráfico das frequências ponderadas da variável uso de antidepressivo.

Pela tabela e gráfico de uso de antidepressivo verifica-se que pouquíssimas pessoas utilizam tais medicamentos, aproximadamente 1,6%. Comparando este dado com o de pessoas que relataram ter depressão (13%) é de reparar que poucos deles usam medicamentos para tratar a doença.

As tabelas a seguir são de avaliação subjetiva sobre como a pessoa se enxerga. As variáveis em análise são mobilidade, cuidado próprio, atividades cotidianas, dor e ansiedade/depressão.

Tabela 17: Frequências da variável mobilidade.

Mobilidade					
Mobilidade	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Não	1686	92.1174	0.8090	90.5211	93.7137
Sim	134	7.8826	0.8090	6.2863	9.4789
Total	1820	100.000			

Tabela 18: Frequências da variável cuidado próprio.

Cuidado próprio					
Cuidado próprio	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Não	1748	95.9784	0.5405	94.9119	97.0449
Sim	72	4.0216	0.5405	2.9551	5.0881
Total	1820	100.000			

Tabela 19: Frequências da variável atividades cotidianas.

Atividades cotidianas					
Atividades cotidianas	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Não	1693	93.1421	0.6695	91.8209	94.4632
Sim	127	6.8579	0.6695	5.5368	8.1791
Total	1820	100.000			

Tabela 20: Frequências da variável dor.

Dor					
Dor	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Não	1129	63.0113	1.7904	59.4787	66.5440
Sim	691	36.9887	1.7904	33.4560	40.5213
Total	1820	100.000			

Tabela 21: Frequências da variável ansiedade/depressão.

Ansiedade/depressão					
Ansiedade/ depressão	**Frequência**	**Percentual**	**Erro Padrão do Percentual**	**Intervalo de Confiança de 95% para o Percentual**	
Não	1399	77.0060	1.1488	74.7392	79.2728
Sim	421	22.9940	1.1488	20.7272	25.2608
Total	1820	100.000			

Pelas tabelas é interessante notar que 37% relatam sentir dores e 23% se dizem ansiosos/depressivos, um número alto considerando que o ideal era não ter nada dessas variáveis.

5.2 Razão de Chances

A tabela a seguir terá cada variável dentro do seu bloco, a razão de chances associada a ela com o respectivo intervalo de confiança e o p-valor relativo à significância da variável em explicar a variável depressão. Para isso foi gerado um modelo diferente para cada variável.

Tabela 22: Tabela com a razão de chances para cada variável.

Variáveis	**Razão de chances**	**IC**		**p-valor**
		Inferior	**Superior**	
1. Bloco socieconômico				

Variáveis	Razão de chances	IC		p-valor
		Inferior	Superior	
1.1 Renda familiar (variável classe_social)				0.4955
Classe D e E	1.209	0.231	6.333	
Classe C	0.993	0.202	4.891	
Classe B	0.786	0.154	4.019	
Classe A2	0.710	0.157	3.209	
Classe A1 (ref)	1			
1.2. Escolaridade (variável p8)				0.0125
Baixa (analfabeto + 1º grau incompleto)	2.079	1.243	3.477	
Média baixa (1º grau completo + 2º grau incompleto)	1.118	0.638	1.957	
Média alta (2º grau complete + 3º grau incompleto)	1.349	0.811	2.243	
Alta (3º grau completo + pós-graduação) (ref)	1			
1.3. Ocupação (variável p9)				0.0528
Não (desempregado + aposentado + não trabalha)	1.395	0.996	1.955	
Sim (servidor + trabalho informal + trabalho doméstico + autônomo) (ref)	1			
2. Bloco demográfico				
2.1. Gênero (variável p4)				<.0001
Feminino	2.030	1.433	2.876	
Masculino (ref)	1			
2.2. Faixa etária (variável fxetaria)				0.0025
56 a 65 anos				
46 a 55 anos				

Variáveis	Razão de chances	IC		p-valor
		Inferior	Superior	
36 a 45 anos	1.856	1.289	2.671	
26 a 35 anos	1.152	0.788	1.685	
18 a 25 anos (ref)				
2.3. Estado conjugal (variável p6)				0.0821
Não-casado (solteiro + separado/divorciado + viúvo)	1.366	0.961	1.942	
Casado (Casado + união consensual) (ref)	1			
2.4. Número de moradores (variável p7)				0.0787
1 a 3 pessoas	1.099	0.704	1.716	
4 a 5 pessoas	0.725	0.474	1.111	
6 em diante (ref)	1			
3. Bloco sobre situação de saúde				
3.1. Diabetes (variável p10)				<.0001
Sim	2.369	1.549	3.623	
Não (ref)	1			
3.2. Hipertensão (variável p11)				<.0001
Sim	3.536	2.431	5.145	
Não (ref)	1			
3.4. Problema cardíaco (variável p13)				<.0001
Sim	6.359	4.023	10.052	
Não (ref)	1			
3.5. Problema respiratório (variável p14)				<.0001
Sim	6.805	4.400	10.526	
Não (ref)	1			

Variáveis	Razão de chances	IC		p-valor
		Inferior	Superior	
3.6. Outros problemas crônicos (variável p15_1)				<.0001
Sim	2.652	1.628	4.319	
Não (ref)	1			
3.7. Plano privado de saúde (variável p16)				0.6385
Sim	1.087	0.769	1.536	
Não (ref)	1			
3.8. Consulta médica (variável p17)				0.0036
Sim	1.610	1.168	2.218	
Não (ref)	1			
3.9. Hospitalização (variável p18)				<.0001
Sim	3.099	2.005	4.788	
Não (ref)	1			
3.10. Uso de antidepressivo (variável AntiDepressivo)				<.0001
Sim	18.675	9.293	37.529	
Não (ref)	1			
4. Bloco de avaliação subjetiva				
4.1. Mobilidade (variável p19)				0.0322
Sim (respostas 2 e 3)	1.825	1.052	3.164	
Não (ref)	1			
4.2. Cuidado próprio (variável p20)				0.0169
Sim (respostas 2 e 3)	2.262	1.158	4.419	
Não (ref)	1			
4.3. Atividades cotidianas (variável p21)				<.0001
Sim (respostas 2 e 3)	4.443	2.864	6.893	

Variáveis	Razão de chances	IC		p-valor
		Inferior	Superior	
Não (ref)	1			
4.4. Dor (variável p22)				<.0001
Sim (respostas 2 e 3)	2.760	1.957	3.893	
Não (ref)	1			
4.5. Ansiedade/depressão (variável p23)				<.0001
Sim (respostas 2 e 3)	7.172	4.803	10.709	
Não (ref)	1			

5.3 Ajuste do Modelo

Uma análise de regressão logística múltipla obedecendo o modelo hierárquico proposto na figura 1, foi ajustado aos dados.

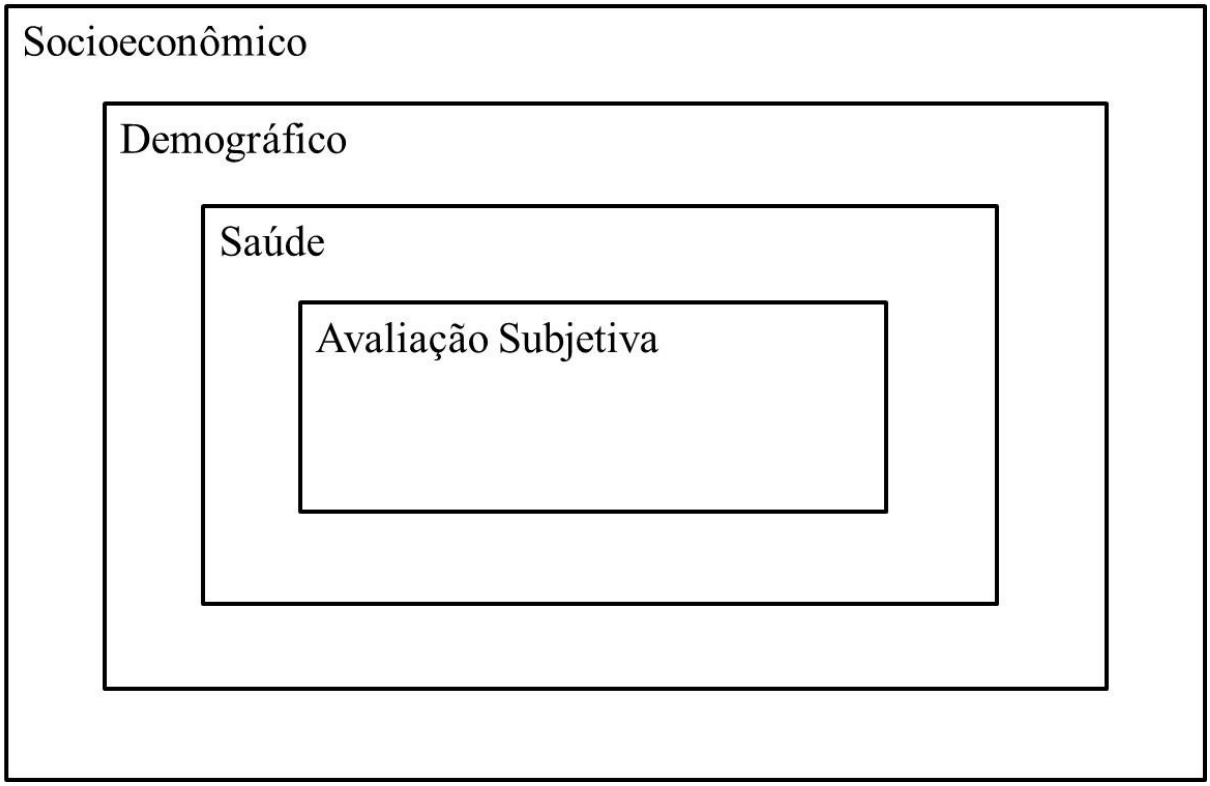

Figura 7 – Estrutura para ajuste do modelo hierárquico.

Os fatores sócio-econômicos, considerados como o principal fator desencadeador de quadros mórbidos e ocorrência de depressão, compõem a primeira etapa de análise. Os fatores demográficos e os relativos a situação de saúde compõem, respectivamente, a segunda e terceira etapas. A quarta etapa de análise é o bloco da avaliação subjetiva.

A inclusão de cada variável dependeu da significância estatística aferida pela razão de chances e seus respectivos intervalos de confiança, sendo incluídas aquelas variáveis que apresentaram $p < 0,1$. Razão de chances ajustadas e os respectivos intervalos de confiança foram obtidos. Para avaliação do modelo será considerado como variável dependente depressão e o procedimento *surveylogistic* será utilizado no SAS para ajustar o modelo. Toda a análise estatística levou em consideração o efeito do delineamento amostral complexo.

Pela saída obtida do SAS abaixo percebe-se que as variáveis classe social e ocupação são retiradas do modelo.

Tabela 23: Análise dos efeitos do modelo gerado.

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Classe social	4	0.5808	0.9652
Escolaridade	3	8.0408	0.0452
Ocupação	1	2.4715	0.1159

O próximo passo é o modelo com as variáveis que passaram na primeira triagem e o bloco demográfico. Com isso têm-se a seguinte tabela:

Tabela 24: Análise dos efeitos do modelo gerado.

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Escolaridade	3	6.2353	0.1007
Gênero	1	15.9455	<.0001
Faixa etária	2	8.2349	0.0163
Estado conjugal	1	1.7331	0.1880
Número de moradores	2	4.3754	0.1122

Com isso, as variáveis estado conjugal e número de moradores saem do modelo e agora o bloco sobre situação de saúde entra na análise.

Tabela 25: Análise dos efeitos do modelo gerado.

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Escolaridade	3	5.9595	0.1136
Gênero	1	4.4173	0.0356
Faixa etária	2	1.2071	0.5469
Diabetes	1	0.0092	0.9238
Hipertensão	1	6.2732	0.0123
Problema cardíaco	1	11.4315	0.0007
Problema respiratório	1	34.7112	<.0001
Outros problemas crônicos	1	1.6778	0.1952
Plano privado de saúde	1	0.0601	0.8063
Consulta médica	1	0.4240	0.5149

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Hospitalização	1	8.6361	0.0033
Uso de antidepressivo	1	39.4368	<.0001

Sobre o bloco de situação de saúde as variáveis diabetes, outros problemas crônicos, plano privado de saúde e consulta médica não se mostraram relevantes para explicar a depressão.

Por fim, para determinação do modelo final em estudo, o bloco de avaliação subjetiva entra na análise. Com isso o seguinte é obtido:

Tabela 26: Análise dos efeitos do modelo gerado.

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Escolaridade	3	4.1681	0.2439
Gênero	1	1.9713	0.1603
Faixa etária	2	1.1543	0.5615
Hipertensão	1	5.9029	0.0151
Problema cardíaco	1	9.8173	0.0017
Problema respiratório	1	33.8401	<.0001
Hospitalização	1	6.2071	0.0127
Uso de antidepressivo	1	32.4239	<.0001
Mobilidade	1	2.2148	0.1367

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Cuidado próprio	1	0.2091	0.6475
Atividades cotidianas	1	8.4899	0.0036
Dor	1	0.5248	0.4688
Ansiedade/ depressão	1	58.2819	<.0001

As variáveis mobilidade, cuidado próprio e dor não são importantes na determinação do modelo.

Para rodar o modelo final foram consideradas as variáveis, na ordem hierárquica, e o resultado foi o seguinte

Tabela 27: Análise dos efeitos do modelo gerado.

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Escolaridade	3	3.4371	0.3290
Gênero	1	2.3069	0.1288
Faixa etária	2	0.9868	0.6105
Hipertensão	1	5.5307	0.0187
Problema cardíaco	1	9.3978	0.0022
Problema respiratório	1	34.5050	<.0001
Hospitalização	1	6.1976	0.0128
Uso de antidepressivo	1	34.1113	<.0001

Análise dos Efeitos			
Efeito	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Atividades cotidianas	1	5.9911	0.0144
Ansiedade/ depressão	1	53.0933	<.0001

Tabela 28: Análise dos coeficientes do modelo gerado.

Análise das Estimativas de Máxima Verossimilhança						
Parâmetro		**DF**	**Estimado**	**Erro Padrão**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.8178	0.3020	159.8068	<.0001
Escolaridade	**Baixa**	1	0.4737	0.3056	2.4034	0.1211
Escolaridade	**Média Alta**	1	0.2580	0.2808	0.8440	0.3582
Escolaridade	**Média Baixa**	1	0.0315	0.3151	0.0100	0.9203
Gênero	**Feminino**	1	0.3105	0.2044	2.3069	0.1288
Faixa etária	**26 - 35**	1	-0.0634	0.2297	0.0762	0.7825
Faixa etária	**36 - 45**	1	0.1675	0.2435	0.4733	0.4915
Hipertensão	**Sim**	1	0.6140	0.2611	5.5307	0.0187
Problema cardíaco	**Sim**	1	1.0339	0.3373	9.3978	0.0022
Problema respiratório	**Sim**	1	1.6131	0.2746	34.5050	<.0001
Hospitalização	**Sim**	1	0.6909	0.2775	6.1976	0.0128
Uso de antidepressivo	**Sim**	1	2.1471	0.3676	34.1113	<.0001

Análise das Estimativas de Máxima Verossimilhança						
Parâmetro		**DF**	**Estimado**	**Erro Padrão**	**Wald Chi-Square**	**Pr > ChiSq**
Atividades cotidianas	**Sim**	1	0.7006	0.2862	5.9911	0.0144
Ansiedade/depressão	**Sim**	1	1.7263	0.2369	53.0933	<.0001

Portanto, o modelo final conta com as variáveis: escolaridade, gênero, faixa etária, hipertensão, problema cardíaco, problema respiratório, hospitalização, uso de antidepressivo, atividades cotidianas e ansiedade/depressão. Os valores dos coeficientes estimados do modelo encontram-se acima para cada variável.

Como o interesse concentra-se na razão de chances, pode-se completar a tabela da seção Razão de Chances com os valores da razão de chances ajustados pelos modelos gerados. Os p-valores também foram substituídos pelos encontrados pelo ajuste hierárquico.

Tabela 29: Razões de chances bruta e ajustada das variáveis.

Variáveis	**Razão de chances bruta**	**IC**		**Razão de chances ajustada**	**IC**		**p-valor**
		Inf	**Sup**		**Inf**	**Sup**	
1. Bloco socieconômico							
1.2. Escolaridade (variável p8)							0.0452
Baixa (analfabeto + 1º grau incompleto)	2.079	1.243	3.477	1.900	1.002	3.605	
Média baixa (1º grau completo + 2º grau incompleto)	1.118	0.638	1.957	1.048	0.528	2.077	
Média alta (2º grau complete + 3º grau incompleto)	1.349	0.811	2.243	1.320	0.754	2.311	
Alta (3º grau completo + pós-graduação) (ref)	1			1			
2. Bloco demográfico							
2.1. Gênero (variável p4)							<.0001

Variáveis	Razão de chances bruta	IC		Razão de chances ajustada	IC		p-valor
		Inf	Sup		Inf	Sup	
Feminino	2.030	1.433	2.876	2.022	1.431	2.857	
Masculino (ref)	1			1			
2.2. Faixa etária (variável fxetaria)							0.0163
56 a 65 anos							
46 a 55 anos							
36 a 45 anos	1.856	1.289	2.671	1.728	1.178	2.535	
26 a 35 anos	1.152	0.788	1.685	1.184	0.799	1.755	
18 a 25 anos (ref)				1			
3. Bloco sobre situação de saúde							
3.2. Hipertensão (variável p11)							0.0187
Sim	3.536	2.431	5.145	1.890	1.143	3.125	
Não (ref)	1			1			
3.4. Problema cardíaco (variável p13)							0.0022
Sim	6.359	4.023	10.052	3.024	1.613	5.668	
Não (ref)	1			1			
3.5. Problema respiratório (variável p14)							<.0001
Sim	6.805	4.400	10.526	4.966	2.924	8.436	
Não (ref)	1			1			
3.9. Hospitalização (variável p18)							0.0033
Sim	3.099	2.005	4.788	2.221	1.290	3.826	
Não (ref)	1			1			

Variáveis	Razão de chances bruta	IC		Razão de chances ajustada	IC		p-valor
		Inf	Sup		Inf	Sup	
3.10. Uso de antidepressivo (variável AntiDepressivo)							<.0001
Sim	18.675	9.293	37.529	15.645	6.568	37.270	
Não (ref)	1			1			
4. Bloco de avaliação subjetiva							
4.3. Atividades cotidianas (variável p21)							0.0036
Sim (respostas 2 e 3)	4.443	2.864	6.893	2.631	1.360	5.090	
Não (ref)	1			1			
4.5. Ansiedade/depressão (variável p23)							<.0001
Sim (respostas 2 e 3)	7.172	4.803	10.709	5.521	3.559	8.566	
Não (ref)	1			1			

Sobre as razões de chance ajustadas das variáveis que se mostraram relevantes no modelo, têm-se que as pessoas com baixa escolaridade possuem 1,9 vezes a mais de chances de ter depressão se comparado com a escolaridade alta; os com escolaridade média baixa praticamente com iguais chances de ser depressivo, com 1,048 vezes mais chances comparado com a escolaridade alta e os de escolaridade média alta com 1,32 vezes mais chances de ter depressão que com de escolaridade alta.

Dentro do bloco demográfico, as mulheres indicaram que têm 2,022 vezes mais chances de serem depressivas. A população entre 36 a 45 anos 1,728 vezes mais chances de apresentarem depressão se comparadas com os de 18 a 25 anos e os de 26 a 35 anos 0,799 vezes mais chances de possuírem depressão que os de 18 a 25 anos.

O bloco sobre situação de saúde mostra as maiores razões de chances encontradas, demonstrando a importância desse bloco no modelo. Os hipertensos indicaram 1,89 vezes de chances de serem depressivos do que os que não são hipertensos; quem possui algum

problema cardíaco 3,024 vezes a mais de chances de serem depressivos do que os que declararam que têm o coração bom. Já os com problemas respiratórios possuem 4,966 vezes mais chances de terem depressão dos que não tem esse problema. As pessoas que já foram hospitalizadas têm 2,221 vezes mais de serem depressivas do que as que nunca foram hospitalizadas. Agora a maior razão de chances do modelo foi a da variável uso de antidepressivos, o que já era de se esperar porque parte-se do pressuposto que a pessoa só usa esse medicamento se é depressiva; então quem usa o remédio tem 15,645 vezes mais chances de serem depressivos.

Apenas duas variáveis do bloco de avaliação subjetiva se mostraram relevantes no modelo. A pessoa que não pratica atividades cotidianas possui 2,631 mais chances de serem depressivas do que os que praticam e os que se sentem ansiosos ou depressivos tem 5,521 vezes mais chances de serem depressivos do que os que não se dizem ansiosos ou depressivos.

Programação Utilizada

Abaixo segue toda a programação SAS utilizada para trabalhar o banco de dados e gerar os resultados.

```
proc import out=dados_final
            datafile="C:\Users\jh\Desktop\Pedro\dados_final.xlsx"
            dbms=xlsx replace;
   sheet='Plan1';
   run;
ods graphics on;
Data dados; set dados_final;
n_domicilios = (3886/182)*(n_domicilios/10);
label p8='Escolaridade'
      p9='Ocupação'
      p4='Gênero'
      fxetaria='Faixa etária'
      classe_social='Classe social'
       p6='Estado conjugal'
       p7='Número de moradores'
       p10='Diabetes'
       p11='Hipertensão'
       p12='Depressão'
       p13='Problema cardíaco'
       p14='Problema respiratório'
       p15_1='Outros problemas crônicos'
       p16='Plano privado de saúde'
       p17='Consulta Médica'
       p18='Hospitalização'
       AntiDepressivo='Uso de antidepressivo'
```

```
       p19='Mobilidade'
       p20='Cuidado próprio'
       p21='Atividades cotidianas'
       p22='Dor'
       p23='Ansiedade/depressão';
run;
data dados1; set dados;
if p8=. or p9=. or p4=. or fxetaria=. or classe_social=. or p6=. or
p7=. or p10=. or
p11=. or p12=. or p13=. or p14=. or p15_1=. or p16=. or p17=. or
p18=. or
AntiDepressivo=. or p19=. or p20=. or p21=. or p22=. or p23=.
then delete;
run;
Proc format;
 value Escolaridade 1='Baixa'
                    2='Baixa'
                           3='Média Baixa'
                           4='Média Baixa'
                           5='Média Alta'
                           6='Média Alta'
                           7='Alta'
                           8='Alta';
 value ocup 1='Sim'
            2='Sim'
                3='Não'
                4='Sim'
                5='Não'
                6='Sim'
                7='Não';
 value conj 1='Não-casado'
            2='Casado'
                3='Não-casado'
                4='Não-casado'
                5='Casado';
 value morad 1-3='1 a 3 pessoas'
             4-5='4 a 5 pessoas'
                6-high='6 em diante';
 value sn 1='Sim'
          2='Não'
             3='Não';
 value sninverso 1='Não'
                  2='Sim'
                     3='Sim';
 value gen 1='Masculino'
           2='Feminino';
 value classe 1='D-E'
              2='C'
                  3='B'
                  4='A2'
                  5='A1';
 value fxetaria 1='18 - 25'
                2='26 - 35'
                      3='36 - 45'
                      4='46 - 55'
```

```
                    5='56 - 65';
run;
ods rtf file="C:\Users\usuario\Desktop\Pedro\word.rtf" ;
proc surveyfreq data=dados;
cluster Setor_cens;
weight n_domicilios;
tables classe_social p8 p9 p4 fxetaria p6 p7 p10 p11 p12 p13 p14
p15_1 p16 p17 p18 AntiDepressivo p19 p20 p21 p22 p23 / cl plots=all;
format classe_social classe.
       p8 Escolaridade.
       p9 ocup.
       p4 gen.
       fxetaria fxetaria.
       p6 conj.
       p7 morad.
       p10 sn.
       p11 sn.
       p12 sn.
       p13 sn.
       p14 sn.
       p15_1 sn.
       p16 sn.
       p17 sn.
       p18 sn.
       AntiDepressivo sn.
       p19 sninverso.
       p20 sninverso.
       p21 sninverso.
       p22 sninverso.
       p23 sninverso.;
run;
ods rtf close;
ods rtf file="C:\Users\jh\Desktop\Pedro\logistic.rtf" ;
proc surveylogistic data=dados1;
format classe_social classe.
       p12 sn.;
class p12 (ref='Não') classe_social (ref='A1');
cluster Setor_cens;
weight n_domicilios;
model p12 = classe_social;
run;
proc surveylogistic data=dados1;
format p8 Escolaridade.
       p12 sn.;
class p12 (ref='Não') p8 (ref='Alta');
cluster Setor_cens;
weight n_domicilios;
model p12 = p8;
run;
proc surveylogistic data=dados1;
format p9 ocup.
       p12 sn.;
class p12 (ref='Não') p9 (ref='Sim');
cluster Setor_cens;
weight n_domicilios;
```

```
model p12 = p9;
run;
proc surveylogistic data=dados1;
format p4 gen.
       p12 sn.;
class p12 (ref='Não') p4 (ref='Masculino');
cluster Setor_cens;
weight n_domicilios;
model p12 = p4;
run;
proc surveylogistic data=dados1;
format fxetaria fxetaria.
       p12 sn.;
class p12 (ref='Não') fxetaria (ref='18 - 25');
cluster Setor_cens;
weight n_domicilios;
model p12 = fxetaria;
run;
proc surveylogistic data=dados1;
format p6 conj.
       p12 sn.;
class p12 (ref='Não') p6 (ref='Casado');
cluster Setor_cens;
weight n_domicilios;
model p12 = p6;
run;
proc surveylogistic data=dados1;
format p7 morad.
       p12 sn.;
class p12 (ref='Não') p7 (ref='6 em diante');
cluster Setor_cens;
weight n_domicilios;
model p12 = p7;
run;
proc surveylogistic data=dados1;
format p10 sn.
       p12 sn.;
class p12 (ref='Não') p10 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p10;
run;
proc surveylogistic data=dados1;
format p11 sn.
       p12 sn.;
class p12 (ref='Não') p11 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p11;
run;
proc surveylogistic data=dados1;
format p13 sn.
       p12 sn.;
class p12 (ref='Não') p13 (ref='Não');
cluster Setor_cens;
```

```
weight n_domicilios;
model p12 = p13;
run;
proc surveylogistic data=dados1;
format p14 sn.
       p12 sn.;
class p12 (ref='Não') p14 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p14;
run;
proc surveylogistic data=dados1;
format p15_1 sn.
       p12 sn.;
class p12 (ref='Não') p15_1 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p15_1;
run;
proc surveylogistic data=dados1;
format p16 sn.
       p12 sn.;
class p12 (ref='Não') p16 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p16;
run;
proc surveylogistic data=dados1;
format p17 sn.
       p12 sn.;
class p12 (ref='Não') p17 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p17;
run;
proc surveylogistic data=dados1;
format p18 sn.
       p12 sn.;
class p12 (ref='Não') p18 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p18;
run;
proc surveylogistic data=dados1;
format AntiDepressivo sn.
       p12 sn.;
class p12 (ref='Não') AntiDepressivo (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = AntiDepressivo;
run;
proc surveylogistic data=dados1;
format p19 sninverso.
       p12 sn.;
class p12 (ref='Não') p19 (ref='Não');
```

```
cluster Setor_cens;
weight n_domicilios;
model p12 = p19;
run;
proc surveylogistic data=dados1;
format p20 sninverso.
       p12 sn.;
class p12 (ref='Não') p20 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p20;
run;
proc surveylogistic data=dados1;
format p21 sninverso.
       p12 sn.;
class p12 (ref='Não') p21 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p21;
run;
proc surveylogistic data=dados1;
format p22 sninverso.
       p12 sn.;
class p12 (ref='Não') p22 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p22;
run;
proc surveylogistic data=dados1;
format p23 sninverso.
       p12 sn.;
class p12 (ref='Não') p23 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p23;
run;
proc surveylogistic data=dados1;
format classe_social classe.
       p8 Escolaridade.
       p9 ocup.
       p12 sn.;
class p12 (ref='Não') classe_social (ref='A1') p8 (ref='Alta') p9
(ref='Sim');
cluster Setor_cens;
weight n_domicilios;
model p12 = classe_social p8 p9;
run;
proc surveylogistic data=dados1;
format p8 Escolaridade.
       p4 gen.
       fxetaria fxetaria.
       p6 conj.
       p7 morad.
       p12 sn.;
```

```
class p12 (ref='Não') p8 (ref='Alta') p4 (ref='Masculino') fxetaria
(ref='18 - 25') p6 (ref='Casado') p7 (ref='6 em diante');
cluster Setor_cens;
weight n_domicilios;
model p12 = p8 p4 fxetaria p6 p7;
run;
proc surveylogistic data=dados1;
format p8 Escolaridade.
       p4 gen.
       fxetaria fxetaria.
       p7 morad.
       p10 sn.
       p11 sn.
       p12 sn.
       p13 sn.
       p14 sn.
       p15_1 sn.
       p16 sn.
       p17 sn.
       p18 sn.
       AntiDepressivo sn.;
class p12 (ref='Não') p8 (ref='Alta') p4 (ref='Masculino') fxetaria
(ref='18 - 25') p7 (ref='6 em diante') p10 (ref='Não')
      p11 (ref='Não') p13 (ref='Não') p14 (ref='Não') p15_1
(ref='Não') p16 (ref='Não') p17 (ref='Não') p18 (ref='Não')
      AntiDepressivo (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p8 p4 fxetaria p7 p10 p11 p13 p14 p15_1 p16 p17 p18
AntiDepressivo;
run;
proc surveylogistic data=dados1;
format p8 Escolaridade.
       p4 gen.
       fxetaria fxetaria.
       p7 morad.
       p11 sn.
       p12 sn.
       p13 sn.
       p14 sn.
       p18 sn.
       AntiDepressivo sn.
       p19 sninverso.
       p20 sninverso.
       p21 sninverso.
       p22 sninverso.
       p23 sninverso.;
class p12 (ref='Não') p8 (ref='Alta') p4 (ref='Masculino') fxetaria
(ref='18 - 25') p7 (ref='6 em diante') p11 (ref='Não')
      p13 (ref='Não') p14 (ref='Não') p18 (ref='Não') AntiDepressivo
(ref='Não') p19 (ref='Não') p20 (ref='Não')
      p21 (ref='Não') p22 (ref='Não') p23 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
```

```
model p12 = p8 p4 fxetaria p7 p11 p13 p14 p18 AntiDepressivo p19 p20
p21 p22 p23;
run;
proc surveylogistic data=dados1;
format p8 Escolaridade.
       p4 gen.
       fxetaria fxetaria.
       p7 morad.
       p11 sn.
       p12 sn.
       p13 sn.
       p14 sn.
       p18 sn.
       AntiDepressivo sn.
       p19 sninverso.
       p20 sninverso.
       p21 sninverso.
       p22 sninverso.
       p23 sninverso.;
class p12 (ref='Não') p8 (ref='Alta') p4 (ref='Masculino') fxetaria
(ref='18 - 25') p7 (ref='6 em diante') p11 (ref='Não')
      p13 (ref='Não') p14 (ref='Não') p18 (ref='Não') AntiDepressivo
(ref='Não') p21 (ref='Não') p23 (ref='Não');
cluster Setor_cens;
weight n_domicilios;
model p12 = p8 p4 fxetaria p7 p11 p13 p14 p18 AntiDepressivo p21
p23;
run;
ods rtf close;
```

6 Referências Bibliográficas

Hauck, W. W., and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 82, 1110-1117.

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*, Second Edition. Wiley, New York.

IBGE. Censo Demográfico 2010, Instituto Brasileiro de Geografia e Estatística.

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, 81, 471-476.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Chapman & Hall, London.

Rao, C. R. (1973). *Linear Statistical Inference and Its Application,* Second Edition. Wiley, Inc., New York.

Silva, P. L. N., Pessoa, D. G. C., e Lila, M. F. (2002). *Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral.* Ciência & Saúde Coletiva, 7 (4): 659-670, 2002.

Roberts, G., Rao, J. N. K., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1-12.

Lehtonen, R., and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys,* Second Edition. John Wiley and Sons, Ltd, England.

Korn, E. L., and Graubard, B. I. (1990). Simultaneous testing of regression coeficients with complex survey data: Use of Bonferroni t statistics, *American Statistician*, **44**, 270-276.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys.* Wiley, Inc., New York.

Thomas, D. R., and Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, **82**, 630-636.